ARL-TR-7294 • MAY 2015

US Army Research Laboratory

Smooth Constrained Heuristic Optimization of a Combinatorial Chemical Space

by Berend Christopher Rinderspacher

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7294 ● MAY 2015

US Army Research Laboratory

Smooth Constrained Heuristic Optimization of a Combinatorial Chemical Space

by Berend Christopher Rinderspacher
Weapons and Materials Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2015	Final	2013–2014

4. TITLE AND SUBTITLE
Smooth Constrained Heuristic Optimization of a Combinatorial Chemical Space

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Berend Christopher Rinderspacher

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-WMM-G
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7294

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
primary author's email: <berend.c.rinderspacher.civ@mail.mil>.

14. ABSTRACT
Several algorithms for optimizing a combinatorial subspace of chemical compound space with constraints are compared. The test system is a library of organic chromophores for electro-optic applications. The hyperpolarizability is maximized while the absorption in the range of 400–700 nm is constrained to zero. The best payoff in terms of primary objective, feasibility, and computational cost is achieved using a heuristic reordering of orthogonal search directions.

15. SUBJECT TERMS
chemical compound space, discrete optimization, nonlinear optics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			B. Christopher Rinderspacher
Unclassified	Unclassified	Unclassified	UU	39	19b. TELEPHONE NUMBER (Include area code) 410-306-2811

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Inverse molecular design (IMD) beyond the purview of drug design has enjoyed increasing popularity recently. Examples can be found in protein design[1–4] or high-hyperpolarizability materials.[5–7] The design problem is complicated by the vastness of possible chemicals, termed chemical space. This space can be viewed as combinatorially complex (e.g., $20^8 \approx 2.6 \cdot 10^{10}$ octapeptides of the naturally occurring amino acids exist alone). As a consequence, a variety of methods have been developed for the discrete optimization in chemical subspaces.[8–15] The continuous optimization of chemicals used in the linear-combination-of-atomic-potentials (LCAP) method[16,17] and the variation-of-particles density-functional-theoretical (VP-DFT) method introduces an important concept for dealing with the inherent roughness of chemical compound space.[18–20] LCAP and VP-DFT interpolate continuously between the Hamiltonians of various chemical species. Furthermore, an investigation into the reasons why chemical optimization is, relatively speaking, "easy" recently utilized probability distributions and expectations on the control variables to arrive at its conclusion.[21,22] Follow-up work discovered that indeed all properties can be optimized efficiently using only the charge density with as-yet unknown functionals.[23] All previous optimizations were executed using at most single constraints. In this contribution, we introduce a flexible method for multiple nonlinear (inequality) constraints applied to a combinatorially complex search space of energetic materials.

In recent years, organic molecules have garnered increasing attention as components of high-hyperpolarizability materials, partly due to the variety of synthetically accessible compounds, cost, and ease of processing.[24,25] Applications for materials with high hyperpolarizabilities are found in telecommunication and optics.[26] The dominant nonlinear response of organic molecules often finds its origin in the conjugated π-system, which facilitates the electronic polarizability. This lends itself to advanced electronic and photonic applications including optical information processing, photovoltaic cells, photodynamic therapy agents, and many others.[26–28] The design of such molecules *in silico* is complicated by the fact that chemical space, even constrained to smaller organic compounds, is combinatorially complex. The number of organic molecules of medium size is estimated to be on the order of 10^{200}.[29] Enumeration is therefore unfeasibly costly and other methods for property optimization need to be developed. Including conformational searching further complicates molecular design.

The optical spectra and nonlinear optical (NLO) properties of chromophores may be conveniently manipulated by modifying the molecular architecture, substitutional groups, or substitution patterns.[26,30–32] The hyperpolarizability tensor are the third derivatives of the energy with respect to the electric field,[33,34]

$$\beta_{ijk} = \frac{\partial^3 E}{\partial F_i \partial F_j \partial F_k}. \tag{1}$$

Commonly only the averaged vector components of the vector, $\beta_i = \frac{1}{3}\sum_j \beta_{ijj} + \beta_{jij} + \beta_{jji}$ are used. In particular the component in the direction of the dipole moment, $\beta_\mu = \frac{1}{||\mu||}\sum_{i=x,y,z} \beta_i \mu_i$, and the vector norm, $\beta_0^2 = \sum_{i=x,y,z} \beta_i^2$, play an important role in electro-optic activity.

Several equivalent methods predict electro-optic trends correctly,[27] despite the abundance of methods and conventions for determining hyperpolarizabilities experimentally.[26,35] The sum-over-states expression of the tensor components in the limit of infinite wavelength,

$$\beta_{ijk} = \sum_{\mu,\nu \neq 0} \frac{\langle 0|x_i|\mu\rangle\langle\mu|x_j|\nu\rangle\langle\nu|x_k|0\rangle}{E_0^\mu E_0^\nu}, \tag{2}$$

where E_0^α is the excitation energy from the ground state to excited state α, links the linear absorption spectrum to the hyperpolarizability.[36] From Eq. 2, low excitation energies coupled with large ground-to-excited-state transition-dipole-moments lead to large hyperpolarizabilities (i.e., a material which absorbs everywhere exhibits maximal hyperpolarizability). But such a material would not be able to transmit anything. Hence, it is necessary to navigate the trade-off between hyperpolarizability and optical constraints by tuning the excited-to-excited-state transition-dipole-moments while maintaining transparency or absorption at the target wavelength(s).

Therefore, the main challenge to designing efficient, electro-optic (EO) materials lies in retaining good linear optical properties. For instance, transparent chromophores in a visible yellow spectral range are typically small molecules with low EO response, while molecules possessing large hyperpolarizabilities are frequently opaque or, at best, have small windows of visibility. The visibility window in these NLO frameworks is generally bracketed by 2 kinds of electronic transitions.[28,30,31] The blue transition is a local excitation of π–π^* character, while

the red transition is characterized by a significant intramolecular charge-transfer. The charge-transfer peak is typically the maximum absorption peak (λ_{max}) while the π-π^* transition has a lower oscillator strength and is often denoted as λ_{max-1}. Optimal chromophores, which are transparent in the visible spectrum, would have a red λ_{max}-peak, while the λ_{max-1}-peak should lie in the blue or UV spectral region to open a large visibility window. In the case of telecommunications applications, it is important to achieve transparency in the 1.30–1.55 μm minimum range.[37] Therefore, the design principle for such chromophores calls for the λ_{max}-peak to be significantly blue-shifted. In this report, we investigate the effects due to the electronic properties of the π-system of the bridge between donor and acceptor.

Computational tools are increasingly playing a role in assisting experimental design of optimal chromophores by curtailing the vast optimization space of chromophore structures.[14,27,38–43] Density-functional theory currently offers the best compromise between accuracy and computational performance for typical chromophores of about 100 atoms. The electronic spectra are often calculated using time-dependent density-functional theory (TD-DFT)[44] and numerous successes of this technique have been recently reviewed.[45]

2. Optimization Methodology

Our optimization problem is formulated as a constrained maximization,

$$\max_{x \in CCS} P(x) \tag{3}$$
$$s.t. : \pi(x) = \sum \epsilon_i^{(x)} \int_{\epsilon_i^{(x)}/700}^{\epsilon_i^{(x)}/400} f_i^{(x)} e^{-\sigma(y-1)^2} dy \leq 30,$$

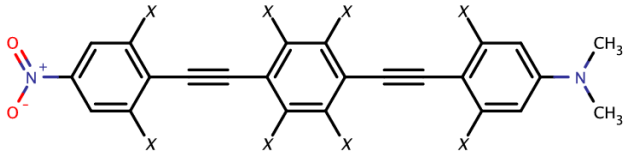
where $\pi(x)$ is the penalty due to the absorbance of compound x from the chosen subspace (CCS) of chemical compound space discussed later in the range 400–700 nm, $\epsilon_i^{(x)}$ and $f_i^{(x)}$ are the excitation wavelength and the normalized oscillator strength of state i for compound x, respectively, $\sigma = 72$ is an approximate broadening coefficient, and $P(x)$ is the hyperpolarizability β_0. The constraint value of 30 is equivalent to an average absorption of 0.1 in the 400- to 700-nm range. We reformulate Problem 3 via a non-negative Lagrange multiplier $0 \leq \lambda \in \mathbb{R}$ for the constraint in the augmented Lagrangian function

$\mathcal{L}(x, \lambda) := P(x) - \lambda\pi(x)$ to

$$\max_{x \in CCS} \min_{\lambda \geq 0} P(x) - \lambda\pi(x). \tag{4}$$

The Problem 4 is then solved alternatingly between the primal variable(s) x and the dual viariables λ.

We begin by describing the optimization procedure within CCS. The chemical subspace that we investigate is depicted in Fig. 1. Each of the 8 Xs on **1** represents a substitution site for which one of 4 chemical groups is attached.



1

Fig. 1 Optimization framework: Each X may be replaced by -H, -F, -Cl, or -Br for a total of 2^{16} possible molecules.

The overall search space thus encompasses 2^{16} molecules. Each such substitution site represents an independent search direction.[14] There is a natural enumeration that follows from the input (see Listing 4 in the Appendix), where the compound $\{x_i = X_i\}$ is assigned the number $n(x) = \sum_i \prod_{j=0}^{i-1} N_j X_i$, where N_j is the number of options on substitution site j and $N_0 = 1$.

The substitutions are enumerated and thus can be viewed as the integer positions on a circle. The optimization proceeds with a local line search for each substitution site in the prescribed order of sites. When neighboring substitutions in the current direction to the current iterate are inferior, the line search is halted and the next direction is searched.

After all search directions have been searched, the minimization with respect to λ is performed using the dual function $P_{(d)}(\lambda) = \max_{x \in CCS} \mathcal{L}(x, \lambda)$ via its approximation

$$P'_{(d)}(\lambda) = \max_{x \in CCS'} \mathcal{L}(x, \lambda) \leq P_{(d)}(\lambda), \tag{5}$$

where CCS' is the subset of visited (i.e., already computed) molecules. Due to the discrete nature of CCS', $P'_{(d)}$ is a piece-wise linear function of λ with general

derivative $-\pi(x(\lambda))$, where $x(\lambda) := \arg\max_{y \in CCS'} \mathcal{L}(y, \lambda)$. Due to the piece-wise linearity, λ only changes meaningfully when $x(\lambda) \neq x(\lambda')$. Thus, whenever $\pi(x) \neq 0$, λ is updated cumulatively by

$$\Delta\lambda = \alpha\pi(x^*) \max\left\{\inf\left\{\lambda' \in \mathbb{R}_+ | \mathcal{L}(x, \lambda^* + \lambda') > \mathcal{L}(x^*, \lambda^* + \lambda')\right\}, 0\right\}, \quad (6)$$

where $\alpha > 1$, CCS' is the set of molecules that have already been computed, x^* is the currently active molecule, and λ^* is the currently active Lagrange multiplier.

By choosing the update along the constraint violation direction $\pi(x^*)$, the update of λ is a steepest descent update with a conservative step size. The step size is chosen such that either x^* violates constraints the least as well as maximizing P or that there exists $x' \in CCS'$ with an improved $\mathcal{L}(x', \lambda + \Delta\lambda) > \mathcal{L}(x^*, \lambda + \Delta\lambda)$. The optimization then returns to optimizing over CCS as described before until no improvement can be attained or the maximum number of steps is exceeded. The general program flow is visualized in Fig. 2.

Since generally there is no *a priori* knowledge of the proper search order of each substituent, the initial assignment is expected to be generally unsuited for smooth optimization. To mitigate this problem, the enumeration of substituents is reassessed after each full cycle of local searches in each direction. For each substituent the average arc tangent of the Lagrangian value is computed,

$$\overline{x}_i^{(j)} = \begin{cases} \frac{\sum_{x \in CCS' | x_i = j} \arctan \mathcal{L}(x,\lambda)}{\#\{x \in CCS' | x_i = j\}}, & \exists x \in CCS' : x_i = j \\ \alpha_i^j, & otherwise \end{cases}, \quad (7)$$

where α_i^j is a default policy for unseen substituents (discussed on the next page), which is subsequently used to order the substituents around the circle starting with the lowest scoring substituent and placing substituents alternatingly to the left and right in ascending order. The resulting order produces a smooth ordering which is monotonically increasing until the maximum is reached, and then monotonically decreasing until the minimum is reached, or vice versa depending on which part of the circle one starts and which direction one goes. The arc tangent is used because some computations may fail due to convergence issues in geometries or general instability of the molecule, which are mapped to a Lagrangian value of $-\infty$.

We investigate 4 different strategies for assigning values to unseen substituents.

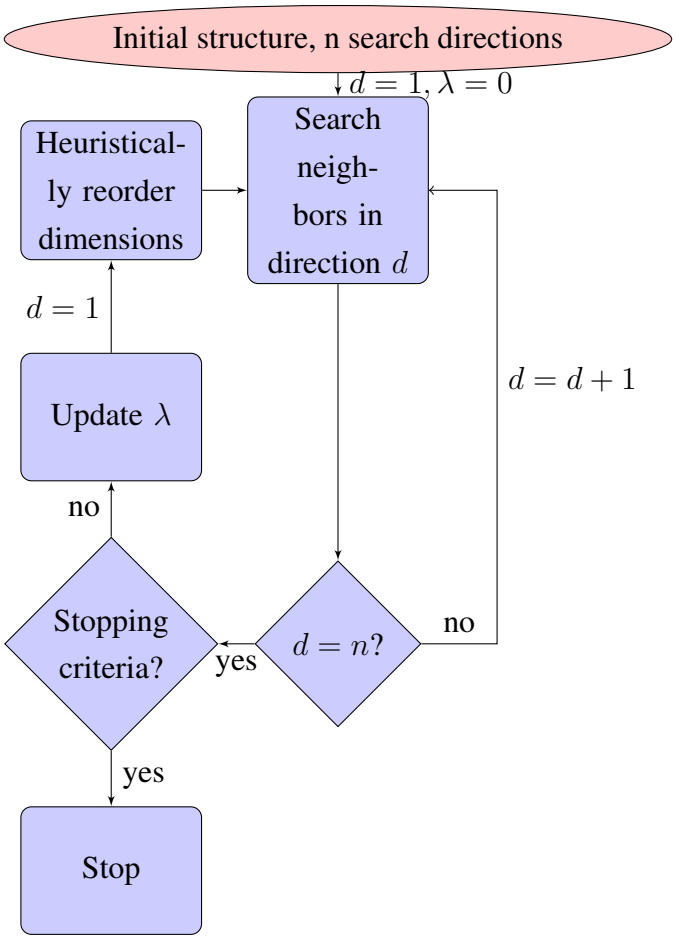


Fig. 2 Flowchart of algorithm

- Algorithm 1: Complete a full sweep of all substituents on the first iteration. After the initial scan there are no unseen substituents anymore.
- Algorithm 2: Assign the unseen substituent the same value as the current iterate; this guarantees that the unseen substituents are immediate neighbors of the current iterate and that at least one will be seen in the next optimization step.
- Algorithm 3: Assign the unseen substituent the value 0. This is easily implemented but may have random results.
- Algorithm 4: Assign the unseen substituents the maximum value. This implies a bias toward unseen substituents to be better in general. As the optimization proceeds this converges with the second method.

No reordering is denoted as Algorithm 0. Four further algorithms are compared:

- Algorithm 5: Conduct a full line search in each direction. No reordering is necessary, but presumably far more compounds need to be computed for each direction.
- Algorithm 6: Reordering occurs as per Algorithm 3, but unlike in Eq. 6, λ is chosen such that $\mathcal{L}(\lambda, x^*) \leq \mathcal{L}(\lambda, x')$, where $x' \in \arg\min_{y \in CCS'} \pi(y)$. Hence, Algorithm 6 solves the subproblem

$$\begin{aligned} &\max_{x \in CCS'} P(x) \\ &s.t. : \pi(x) \leq 30. \end{aligned} \tag{8}$$

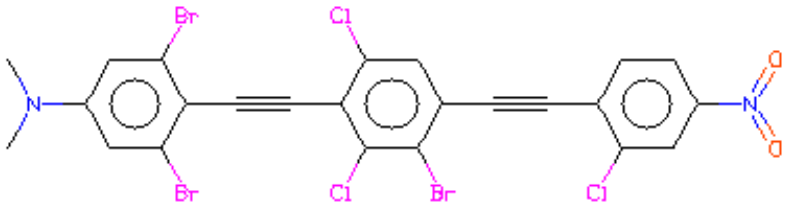
- Algorithm 7: Following the protocol of optimizing on a hypercube.[14]
- Algorithm 8: Same as Algorithm 7, but with deleting compounds x with $\pi(x) > \pi(x^*)$ from the library, effectively reordering the library, where x^* is the current iterate.

Since the latter 2 methods are unaware of the underlying structure of the search space, they may be expected to perform erratically.

3. Computational Chemistry Protocol

All geometries were optimized with PM6 using *Gaussian 09*[46] as described elsewhere[47] (see Listing 1 in the Appendix for specifics). Hyperpolarizabilities β_μ were computed at the CNDO level of theory.[14] Spectra were computed at the BH&HLYP/6-31+G(d) level in the time-dependent density-functional framework[48,49] using *Gaussian 09*.[46]

Before further analysis, the conformational space was explored to ensure only low-energy species were considered.[47] Listing 2 in the Appendix summarizes these steps.



2

Fig. 3 Best candidate found

4. Results and Discussion

Despite expectations Algorithm 5 does not always perform better at reducing the violation than the other algorithms (see start number 41668 in Table 1). As expected, the Algorithm 7 variants tend to perform worse than the Algorithm 5 variants. Surprisingly, the Algorithm 8 actually performs worse on average than Algorithm 7. Algorithm 6 uses the fewest iterations on average per individual run. The largest number of computed compounds in a single run was 138, which constitutes merely 0.02% of the library. Despite minimizing π, Algorithm 6 results in the lowest penalty violation once (see start number 41668 in Table 1). Except for start number 8389, Algorithm 0 is inferior with respect to fulfilling constraints to all other methods but Algorithm 8, which fails to outperform Algorithm 0 for start number 0 only.

Table 1 Comparison of the performance of the assorted reordering schemes on multiple starting compounds

Method	Start	# Molecules	Property	Penalty	Efficiency
Algorithm 5	0	119	121.31	15.5295	0.000541122
Algorithm 0		85	133.88	23.2389	0.000506251
Algorithm 1		81	124.352	21.5181	0.000573735
Algorithm 2		74	133.88	23.2389	0.000581504
Algorithm 3		74	124.352	21.5181	0.000628007
Algorithm 4		93	126.962	18.8063	0.00057176
Algorithm 6		49	124.352	21.5181	0.000948418
Algorithm 7		64	132.119	23.5097	0.000664619
Algorithm 8		80	130.764	27.0287	0.000462471
Algorithm 5	4710	114	121.31	15.5295	0.000564856

Table 1 Comparison of the performance of the assorted reordering schemes on multiple starting compounds (continued)

Method	Start	# Molecules	Property	Penalty	Efficiency
Algorithm 0		45	164.923	38.6927	0.000574326
Algorithm 1		84	139.494	21.5761	0.000551757
Algorithm 2		91	121.31	15.5295	0.000707622
Algorithm 3		93	121.31	15.5295	0.000692404
Algorithm 4		93	121.31	15.5295	0.000692404
Algorithm 6		76	125.011	17.7595	0.000740893
Algorithm 7		55	125.011	17.7595	0.00102378
Algorithm 8		48	163.287	35.6982	0.000583596
Algorithm 5	8389	120	105.054	3.5658	0.002337016
Algorithm 0		111	105.054	3.5658	0.002526504
Algorithm 1		118	105.054	3.5658	0.002376627
Algorithm 2		139	105.339	8.68666	0.000828195
Algorithm 3		139	105.333	8.68666	0.000828195
Algorithm 4		128	125.799	18.8994	0.000413373
Algorithm 6		63	93.8483	5.66247	0.002803196
Algorithm 7		64	104.418	8.92695	0.001750318
Algorithm 8		81	105.333	8.68666	0.001421223
Algorithm 5	41329	142	105.054	3.5658	0.001974943
Algorithm 0		75	164.923	38.6927	0.000344596
Algorithm 1		111	112.876	11.5523	0.000779845
Algorithm 2		111	111.944	7.44892	0.001209438
Algorithm 3		112	105.054	3.5658	0.002503946
Algorithm 4		111	111.944	7.44892	0.001209438
Algorithm 6		85	96.982	5.35516	0.002196892
Algorithm 7		77	129.505	23.4977	0.000552693
Algorithm 8		97	125.799	18.8994	0.000545482
Algorithm 5	41668	107	115.973	12.729	0.000734213
Algorithm 0		88	132.798	22.3781	0.000507802
Algorithm 1		104	93.8483	5.66247	0.00169809
Algorithm 2		137	126.962	18.8063	0.000388129
Algorithm 3		87	115.973	12.729	0.000902997
Algorithm 4		138	126.962	18.8063	0.000385316

Table 1 Comparison of the performance of the assorted reordering schemes on multiple starting compounds (continued)

Method	Start	# Molecules	Property	Penalty	Efficiency
Algorithm 6		101	105.054	3.5658	0.002776653
Algorithm 7		87	118.986	13.5325	0.000849381
Algorithm 8		112	105.333	8.68666	0.001037109

Table 2 Ratio of objective to constraint, rounded to uniquely identifying ratios

Method	0	4710	8389	41329	41668
Algorithm 5	7.81	7.81	29.46	29.46	9.11
Algorithm 0	5.76	4.26	29.46	4.26	5.93
Algorithm 1	5.78	6.47	29.46	9.77	16.57
Algorithm 2	5.76	7.81	12.13	15.03	6.75
Algorithm 3	5.78	7.81	12.13	29.46	9.11
Algorithm 4	6.75	7.81	6.66	15.03	6.75
Algorithm 6	5.78	7.04	16.57	18.11	29.46
Algorithm 7	5.62	7.04	11.66	5.51	8.79
Algorithm 8	4.84	4.57	12.13	6.66	12.13

Table 3 Comparison of average performance

Method	Total # Molecules	Total Efficiency	Average Efficiency
Algorithm 5	602	0.00046585	0.00123043
Algorithm 0	404	0.000694163	0.000891896
Algorithm 1	498	0.000563136	0.001196011
Algorithm 2	552	0.000243202	0.000742978
Algorithm 3	505	0.000555331	0.00111111
Algorithm 4	563	0.000238451	0.000654458
Algorithm 6	374	0.000749845	0.00189321
Algorithm 7	347	0.000322825	0.000968158
Algorithm 8	418	0.000275404	0.000809976

5. Conclusions

We have developed and compared efficient algorithms for handling constraints in the optimization of substitutional subspaces of chemical compound space. The best candidate was found to be **2** (see Fig. 3). Algorithm 6 shows the most promise for finding good candidates with minimal effort both for single runs as well as aggregate samples. Reordering substitutions improves expected performance for single runs as both Algorithm 1 and Algorithm 3 are very competitive with Algorithm 6. Since Algorithm 6 differs from Algorithm 3 merely in solving the subproblem, there is no additional computational cost associated. Algorithm 4 shows the worst performance both in aggregate as on average due to relegating entire subspaces maximally far from the current iterate.

The Algorithm 7 family of algorithms performs surprisingly well on average, even though it does not utilize the underlying structure of the search space. On the other hand, aggregate efficiency is poor for the Algorithm 7 family.

Since Algorithm 7 consistently computes the fewest compounds before converging, it can be recommended when only a quick assessment is required. In all other cases, Algorithm 6 should be considered superior.

6. References

1. Bolon DN, Mayo SL. Enzyme-like proteins by computational design. Proc Nat Acad Sci USA. 2001;98(25):14274–14279.

2. Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. Nat Struct Biol. 2003;10(1):45–52.

3. Park S, Xi Y, Saven JG. Advances in computational protein design. Curr Opin Struct Biol. 2004;14(4):487–494.

4. Mang NG, Zeng C. Reference energy extremal optimization: A stochastic search algorithm applied to computational protein design. J Comp Chem. 2008;29(11):1762–1771.

5. Dudiy SV, Zunger A. Searching for alloy configurations with target physical properties: Impurity design via a genetic algorithm inverse band structure approach. Phys Rev Lett. 2006;97(4).

6. Franceschetti A, Dudiy SV, Barabash SV, Zunger A, Xu J, van Schilfgaarde M. First-principles combinatorial design of transition temperatures in multicomponent systems: The case of mn in gaas. Phys Rev Lett. 2006;97(4).

7. Piquini P, Graf PA, Zunger A. Band-gap design of quaternary (in,ga)(as,sb) semiconductors via the inverse-band-structure approach. Phys Rev Lett. 2008;100(18).

8. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671–680.

9. Bohachevsky IO, Johnson ME, Stein ML. Generalized simulated annealing for function optimization. Technometrics. 1986;28(3):209–217.

10. Muhlenbein H, Gorgesschleuter M, Kramer O. Evolution algorithms in combinatorial optimization. Par Comput. 1988;7(1):65–85.

11. Desmet J, Demaeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature. 1992;356(6369):539–542.

12. Allen BD, Mayo SL. Dramatic performance enhancements for the faster optimization algorithm. J Comp Chem. 2006;27(10):1071–1075.

13. Keinan S, Paquette WD, Skoko JJ, Beratan DN, Yang WT, Shinde S, Johnston PA, Lazo JS, Wipf P. Computational design, synthesis and biological evaluation of para-quinone-based inhibitors for redox regulation of the dual-specificity phosphatase cdc25b. Org Biom Chem. 2008;6(18):3256–3263.

14. Rinderspacher BC, Andzelm J, Rawlett A, Dougherty J, Beratan DN, Yang W. Discrete optimization of electronic hyperpolarizabilities in a chemical subspace. J Chem Theory Comput. 2009;5(12):3321–3329.

15. Weymuth T, Reiher M. Gradient-driven molecule construction: An inverse approach applied to the design of small-molecule fixating catalysts. Int J Quantum Chem. 2014;114(13):838–850.

16. Wang ML, Hu XQ, Beratan DN, Yang WT. Designing molecules by optimizing potentials. J Am Chem Soc. 2006;128(10):3228–3232.

17. Keinan S, Hu XQ, Beratan DN, Yang WT. Designing molecules with optimal properties using the linear combination of atomic potentials approach in an AM1 semiempirical framework. J Phys Chem A. 2007;111(1):176–181.

18. von Lilienfeld OA, Lins RD, Rothlisberger U. Variational particle number approach for rational compound design. Phys Rev Lett. 2005;95(15).

19. von Lilienfeld OA, Tavernelli I, Rothlisberger U, Sebastiani D. Variational optimization of effective atom centered potentials for molecular properties. J Chem Phys. 2005;122(1).

20. von Lilienfeld OA, Tuckerman ME. Molecular grand-canonical ensemble density functional theory and exploration of chemical space. J Chem Phys. 2006;125(15).

21. Moore KW, Pechen A, Feng XJ, Dominy J, Beltrani VJ, Rabitz H. Why is chemical synthesis and property optimization easier than expected? Phys Chem Chem Phys. 2011;13:10048–10070.

22. Green DF. A Statistical Framework for Hierarchical Methods in Molecular Simulation and Design. J Chem Theory Comput. 2010;6(5):1682–1697.

23. Rinderspacher BC. Chemical compound design using nuclear charge distributions. Army Research Laboratory (US); 2012. Report No.: ARL-TR-5935.

24. Andrekson PA, Westlund M. Nonlinear optical fiber based high resolution all-optical waveform sampling. Las Phot Rev. 2007;1(3):231–248.

25. Bergmann G, Ellis C, Hindmarsh P, Kelly SM, O'Neill M. The synthesis and characterisation of nematic nlo polymer networks. Mol Cryst and Liq Cryst. 2001;368:4439–4446.

26. Dalton LR, Sullivan PA, Bale DH. Electric field poled organic electro-optic materials: State of the art and future prospects. Chem Rev. 2010;110(1):25–55; PMID: 19848381.

27. Isborn CM, Leclercq A, Vila FD, Dalton LR, Bredas JL, Eichinger BE, Robinson BH. Comparison of static first hyperpolarizabilities calculated with various quantum mechanical methods. J Chem Phys A. 2007;111(7):1319–1327.

28. Cheng YJ, Luo J, Hau S, Bale DH, Kim TD, Shi Z, Lao DB, Tucker NM, Tian Y, Dalton LR, Reid PJ, Jen AKY. Large electro-optic activity and enhanced thermal stability from diarylaminophenyl-containing high-β nonlinear optical chromophores. Chem Mat. 2007;19(5):1154–1163.

29. van Deursen R, Reymond JL. Chemical space travel. Chem Med Chem. 2007;2:636–640.

30. Chidthong R, Hannongbua S. Excited state properties, fluorescence energies, and lifetimes of a poly(fluorene-phenylene), based on TD-DFT Investigation. J Comp Chem. 2010;31(7):1450–1457.

31. Jansson E, Minaev B, Schrader S, Agren H. Time-dependent density functional calculations of phosphorescence parameters for fac-tris(2-phenylpyridine) iridium. Chem Phys. 2007;333(2-3):157–167.

32. Nasr R, Hirschberg DS, Baldi P. Hashing algorithms and data structures for rapid searches of fingerprint vectors. J Chem Inf Mod. 2010;50(8):1358–1368.

33. d'Avezac M, Botts R, Mohlenkamp MJ, Zunger A. learning to predict physical properties using sumes of separable functions. SIAM J Sci Comput. 2011;33(6):3381–3401.

34. Champagne B, Perpete E, Jacquemin D, van Gisbergen S, Baerends E, Soubra-Ghaoui C, Robins K, Kirtman B. Assessment of conventional density functional schemes for computing the dipole moment and (hyper)polarizabilities of push-pull pi-conjugated systems. J Phys Chem A. 2000;104(20):4755–4763.

35. Willetts A, Rice JE, Burland DM, Shelton DP. Problems in the comparison of theoretical and experimental hyperpolarizabilities. J Chem Phys. 1992;97(10).

36. Obare SO, De C, Guo W, Haywood TL, Samuels TA, Adams CP, Masika NO, Murray DH, Anderson GA, Campbell K, Fletcher K. Fluorescent chemosensors for toxic organophosphorus pesticides: A review. Sensors. 2010;10(7):7018–7043.

37. Baldi P, Nasr R. When is chemical cimilarity significant? The statistical distribution of chemical similarity scores and its extreme values. J Chem Inf Mod. 2010;50(7):1205–1222.

38. Schmidt E. Zur Theorie der linearen und nichtlinearen Integralgleichungen [On the theory of linear and nonlinear integral equations]. Math Annalen. 1907;63:443.

39. Suramitr S, Hannongbua S, Wolschann P. Conformational analysis and electronic transition of carbazole-based oligomers as explained by density functional theory. J Mol Struc–Theochem. 2007;807(1-3):109–119.

40. Zerner M, Reidlinger C, Fabian W, Junek H. Push-pull dyes containing malononitrile dimer as acceptor: synthesis, spectroscopy and quantum chemical calculations. J Mol Struc–Theochem. 2001;543:129–146.

41. Poprawa-Smoluch M, Baggerman J, Zhang H, Maas HPA, De Cola L, Brouwer AM. Photoisomerization of Disperse Red 1 studied with transient absorption spectroscopy and quantum chemical calculations. J Phys Chem A. 2006;110(43):11926–11937.

42. Andzelm J, Rinderspacher BC, Rawlett A, Dougherty J, Baer R, Govind N. Performance of DFT methods in the calculation of optical spectra of TCF-chromophores. J Chem Theory Comput. 2009;5(10):2835–2846.

43. Breitung E, Shu C, McMahon R. Thiazole and thiophene analogues of donor-acceptor stilbenes: Molecular hyperpolarizabilities and structure-property relationships. J Am Chem Soc. 2000;122(6):1154–1160.

44. Tokarski J, Hopfinger A, Hobbs J, Ford D, Faulon J. Molecular modelling of polymers 17. Simulation and QSPR analyses of transport behavior in amorphous polymeric materials. Comput Theor Polym S. 1997;7(3-4, 1):199–214.

45. Faulon J, Visco D, Pophale R. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. J Chem Inf Comput Sci. 2003;43(3):707–720.

46. Gaussian 09. Wallingford (CT): Gaussian Inc. 2004 [accessed 2014 Dec 23]. http://www.gaussian.com.

47. Rinderspacher BC. Discrete optimization in chemical space reference manual. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2012 Oct. Report No.: ARL-TR-6202.

48. Rinderspacher BC, Andzelm JW, Rawlett AM, Dougherty JM, Baranoski M, Davis MC. The role of aromatic π-bridges in push-pull-chromophores on the transparency-hyperpolarizability tradeoff. Chem Phys Lett. 2011;507(4-6):221–225.

49. Rinderspacher BC. Electro-optic and spectroscopic properties of push-pull-chromophores with non-aromatic π-bridges. Chem Phys Lett. 2013;585(0):21–26.

Appendix. Listings

Listing 1 energy_run.sh runs the constrained geometry preoptimization and the follow-up full optimization. It returns the final coordinates.

```
#!/bin/bash
# Execution script
# remove extra files
EXEC=$PWD/g09_run
filename=`basename ${1} .dat`
mkdir -p $1

if [ ! -e "$1/energy" -a ! -e "$1/failed" ]; then
    cd $1
    echo %chk=opt.chk > pre.com
    cat ../header.com >> pre.com
    cat ../$1.zmat >> pre.com
    echo %chk=opt.chk > opt.com
    cat ../footer.com >> opt.com
    if [ ! -e "pre.log" ]; then
       NORMALEXEC=`tail pre.log | grep -o termination `
       if [ ! -z "$NORMALEXEC" -o ! -e "opt.chk" ]; then
          $EXEC pre.com pre.log
       fi
    fi
    $EXEC opt.com opt.g09_out
    NORMALEXEC=`tail opt.g09_out | grep -o Normal| awk '{print\
      $1}'`
    if [ "$NORMALEXEC" == "Normal" ]; then
        fgrep opt.g09_out -e 'SCF Done' | tail -1 | awk '{print\
          $5}' > energy
        awk -f ../logcart.awk < opt.g09_out > xyz
        python ~/bin/retrieve_zmat_from_xyz.py ../$1.zmat xyz \
           | grep c > opt.rconsts
        python ~/bin/retrieve_zmat_from_xyz.py ../$1.zmat xyz \
           | grep d > opt.rvars
        ../retrieve_zmat_g03 opt.g09_out opt
```

This appendix appears in its original form, without editorial changes.

```
        bzip2 opt.g09_out
        bzip2 *.chk
        cd ..
        ln -s $1/energy $1.energy
        ln -s $1/opt.rvars $1.rvars
        ln -s $1/opt.rconsts $1.rconsts
    else
        echo -1000000 > result
        touch failed
        exit 1
    fi
fi
```

Listing 2 proprty_script.sh computes the properties and penalties of a molecule.

```
#!/bin/bash
# Execution script
# geometry optimize in Gaussian and feed to CNDO
# run cndo
# read strongest excitation
# read beta
# remove extra files
EXEC=/apps/gaussian/scripts/g09l_run_d01
filename=`basename ${1} .dat`
mkdir -p $1

if [ ! -e "$1/$1.failed" -a -e "$1/$1.xyz" ]; then
    cd $1
# Compute Hyperpolarizability
        if [ ! -e "$1.result" ];  then
# CNDO
        ../xyz_to_CNDO $1.xyz $1.dat 0 1
        ln -s ../SOS_input.txt .
        ln -s ../INDO1S.par .
        echo $1 | cndo >& cndo_out
        SOSx
        ../clean_cndo
```

```
        mv sos* $1.sos
        fgrep $1.sos -e 'beta mu' | awk '{print $3}' > $1.result
        BLA=`cat $1.result`
        ../fabs $BLA > $1.result
        bzip2 $1.log
        cd ..
        ln -s $1/$1.result .
        cd $1
        fi
        if [ ! -e "$1.penalty" ]; then
        bunzip2 $1.chk.bz2

# Compute penalty of spectrum

        echo %chk=$1.chk > $1_spectrum.com
        cat ../spectrum.com >> $1_spectrum.com
        $EXEC $1_spectrum.com $1_spectrum.log 32
        bzip2 $1.chk
        STATEG09=`fgrep $1_spectrum.log -e Normal | awk \
'{print $1}'`
        if [ "$STATEG09" == "Normal" ]; then
            fgrep $1_spectrum.log -e "Excited␣State␣" | \
awk '{print $7}' > $1.nm
            fgrep $1_spectrum.log -e "Excited␣State␣" | \
awk '{print $9}' | awk -F = '{print $2}' > $1.osc
            ../integrate_window_penalty ${1}.nm $1.osc \
${1}.penalty
            BLA=`cat $1.penalty`
            ../fabs $BLA > $1.penalty

            rm $1.chk.bz2
            bzip2 *log
        else
            echo 100000 > $1.penalty
                touch $1.failed
            exit 1
```

```
        fi
        cd ..
        ln -s $1/$1.penalty .
        exit 0
        fi
fi
```

Listing 3 Script that sets up the right links for the optimization after a good initial structure has been identified.

```
#!/bin/bash

n=$1
echo $n
ln -s $1 $2
for i in `ls  ${n}.* | awk -F "${n}" '{print $2}'`; do
#       echo $i
        if [ ! -e "$2$i" ]; then
        ln -s $1$i $2$i
        fi
done
n=$1
cd $1
for i in `ls ${n}*chk* | awk -F "$n" '{print $2}'`; do
        cp $1$i $2$i
done
for i in `ls ${n}*.* | awk -F "$n" '{print $2}'`; do
#       echo $i
        if [ ! -e "$2$i" ]; then
        ln -s $1$i $2$i
        fi
done
cd ..
```

Listing 4 Input file to determine the substitutional search space.

```
ChemGroup((
Z( # Framework
```

```
 (N ,−3 , 0.000000 , −2 , 0.000000 ,−1 , 0.000000
)
 (C , 0 , 1.455198 , −3 , 0.000000 , −2, 0.000000
)
 (C , 0 , 1.455202 , 1 ,118.635000 ,−3 , 0.000000
)
 (C , 0 , 1.382367 , 1 ,119.869000 , 2 , 165.442000
)
 (H , 1 , 1.090564 , 0 ,109.068000 , 3 ,−177.996000
)
 (H , 1 , 1.100788 , 0 ,112.488000 , 3 , 61.924000
)
 (H , 1 , 1.096133 , 0 ,111.145000 , 3 , −59.556000
)
 (H , 2 , 1.090564 , 0 ,109.068000 , 3 , 178.005000
)
 (H , 2 , 1.100785 , 0 ,112.487000 , 3 , −61.916000
)
 (H , 2 , 1.096133 , 0 ,111.145000 , 3 , 59.565000
)
 (C , 3 , 1.418413 , 0 ,121.400000 , 2 , 172.824000
)
 (C ,10 , 1.387462 , 3 ,121.166000 , 0 , 178.587000
)
 (C ,11 , 1.410705 , 10 ,121.553000 , 3 , 0.387000
)
 (C ,12 , 1.410706 , 11 ,117.355000 ,10 , 0.288000
)
 (C ,13 , 1.387460 , 12 ,121.552000 ,11 , −0.288000
)
 (C ,12 , 4.060996 , 11 ,121.324000 ,10 ,−179.837000
)
 (C ,15 , 1.413713 , 12 ,120.840000 ,11 , 0.091000
)
 (C ,16 , 1.387454 , 15 ,120.846000 ,12 , 179.824000
)
```

(C ,17 , 1.412401 , 16 ,120.748000 ,15 , 0.004000
)
(C ,18 , 1.412399 , 17 ,118.500000 ,16 , 0.022000
)
(C ,15 , 1.413711 , 16 ,118.313000 ,17 , −0.030000
)
(C ,18 , 4.061134 , 17 ,120.750000 ,16 ,−179.910000
)
(C ,21 , 1.413807 , 18 ,120.572000 ,17 , −0.003000
)
(C ,22 , 1.388815 , 21 ,120.703000 ,18 , 179.943000
)
(C ,23 , 1.396983 , 22 ,118.985000 ,21 , 0.002000
)
(C ,24 , 1.396982 , 23 ,121.768000 ,22 , 0.002000
)
(C ,21 , 1.413807 , 22 ,118.856000 ,23 , −0.006000
)
(N ,24 , 1.467005 , 23 ,119.116000 ,22 ,−179.996000
)
(O ,27 , 1.233931 , 24 ,117.856000 ,23 , 0.008000
)
(O ,27 , 1.233930 , 24 ,117.855000 ,23 ,−179.992000
)
(H ,23 , 1.082963 , 22 ,121.410000 ,21 ,−179.995000
)
(H ,25 , 1.082964 , 24 ,119.605000 ,23 , 180.000000
)
(H ,10 , 1.082697 , 11 ,118.420000 ,12 ,−179.648000
)
(H ,14 , 1.082697 , 13 ,118.420000 ,12 , 179.647000
)
(C ,12 , 1.420584 , 11 ,121.322000 ,10 ,−179.945000
)
(C ,15 , 1.420221 , 16 ,120.843000 ,17 , 179.930000
)

```
 (C ,18 , 1.421365 , 17 ,120.750000 ,16 ,-179.945000
)
 (C ,21 , 1.420438 , 22 ,120.572000 ,23 , 179.981000
)
)
ReturnConnector()
Connector(
 (
  (16,15,35)
  (0.7, 0, 0)
  (0,120,0)
  (0,0,0)
  ( 0, 0, 1)
  ( 0, 0, 1)
  ( 0, 0, 0)
  ()
  )
  (
   (20,15,35)
   (0.7, 0,0)
   (0,120,0)
   (0,0,0)
  ( 0, 0, 1)
  ( 0, 0, 1)
  ( 0, 0, 0)
   ()
  )
  (
   (13,12,11)
   (0.7, 0,0)
   (0,120,0)
   (0,0,180)
  ( 0, 0, 1)
  ( 0, 0, 1)
  ( 0, 0, 0)
   ()
```

```
)
(
 (11,12,13)
 (0.7, 0,0)
 (0,120,0)
 (0,0,180)
( 0, 0, 1)
( 0, 0, 1)
( 0, 0, 0)
 ()
)
(
 (17,16,15)
 (0.7, 0,0)
 (0,120,0)
 (0,0,180)
( 0, 0, 1)
( 0, 0, 1)
( 0, 0, 0)
 ()
)
(
 (19,18,17)
 (0.7, 0,0)
 (0,120,0)
 (0,0,180)
( 0, 0, 1)
( 0, 0, 1)
( 0, 0, 0)
 ()
)
(
 (22,21,37)
 (0.7, 0,0)
 (0,120,0)
 (0,0,0)
```

```
    ( 0, 0, 1)
    ( 0, 0, 1)
    ( 0, 0, 0)
     ()
    )
    (
     (26,21,37)
     (0.7, 0,0)
     (0,120,0)
     (0,0,0)
    ( 0, 0, 1)
    ( 0, 0, 1)
    ( 0, 0, 0)
     ()
    )
  )
allowed_groups(
   (1,2,3,4) # Subs, H, F, Cl, Br
   (1,2,3,4)
   (1,2,3,4)
   (1,2,3,4)
   (1,2,3,4)
   (1,2,3,4)
   (1,2,3,4)
   (1,2,3,4)
  )
)

   (#1
    Z(
     (H, -3, 0.45, -2, 0, -1, 0)
    )
    ReturnConnector()
    Connector()
    allowed_groups()
   )
```

```
   (#2
    Z(
     (F, -3, 0.6, -2, 0, -1, 0)
    )
    ReturnConnector()
    Connector()
    allowed_groups()
   )
   (#3
    Z(
     (Cl, -3, 0.8, -2, 0, -1, 0)
    )
    ReturnConnector()
    Connector()
    allowed_groups()
   )
   (#4
    Z(
     (Br, -3, 1.2, -2, 0, -1, 0)
    )
    ReturnConnector()
    Connector()
    allowed_groups()
   )
)
nconstraints(1)
```

Listing 5 integrate_window_penalty.cc Computes the penalty function for absorbing in the visible electromagnetic range.

```
#include <fstream>
#include <iostream>
#include <sstream>
#include <string>
#include <cmath>

#define MAX_STEP 20
```

```
using namespace std;

double tolerance=1e-14;

double num_int(double x, double y,double alpha)
{
    double d=y-x;
    double dx=d/100000.0;
    double r=0.0;
    for(long i=0;i<100001;i++,x+=dx)
          r+=dx*exp(-(1.0-x)*(1.0-x)/(x*x)*alpha*alpha);
    return r;
}

double diff_int_Gaussian(double x,double y)
{
  long i;
  double z=1;
  double r=0.0;
  const double x2=x*x;
  const double x4=x2*x2;
  const double y2=y*y;
  const double y4=y2*y2;
  double x4i=x;
  double y4i=y;
  double r1=0.0;
  double r2=0.0;

  for(i=0;i<MAX_STEP && z>tolerance; i++)
    {
      z=1.0/(double) (4*i+1)*(y4i-x4i)-
        (y2*y4i-x2*x4i)/(double) ((4*i+3)*(2*i+1));
      r1+=1.0/(double) (4*i+1)*(x4i)-
        (x2*x4i)/(double) ((4*i+3)*(2*i+1));
      r2+=1.0/(double) (4*i+1)*(y4i)-
        (y2*y4i)/(double) ((4*i+3)*(2*i+1));
```

```
      r+=z;
      x4i*=x4/(double) ((2*i+1)*(2*i+2));
      y4i*=y4/(double) ((2*i+1)*(2*i+2));
    }
  return r2-r1;
}


int main(int argc, char *argv[])
{
  if (argc <3)
    cerr << "Usage: " << argv[0]
         << " <nmfile> <oscfile> <NLO_optresultfile>\n";
  ifstream log(argv[1]);
  ifstream log2(argv[2]);
  fstream result;

  bool done=false;
  string search;

  int number;
  double eV,cm_1,nm;
  double strength;
  double penalty=-30.0;
    log >> nm;
    log2 >> strength;
    double alpha=sqrt(72.0);
  while(log.good() && !done) {
    penalty+=strength*num_int(400.0/nm,700.0/nm,alpha)*nm;
    log >> nm;
    log2 >> strength;
  }
  if(penalty<0) penalty=0.0;
  result.open(argv[3],ios::in);
  getline(result,search);
```

```
    result.close();
    stringstream ss2;
    double val=0.0;
    ss2 << search;
    ss2 >> val;
    val-=penalty;
    result.open(argv[3],ios::out);
    result << fabs(val) << endl;
    result.flush();
    result.close();
}
```

List of Symbols, Abbreviations, and Acronyms

CCS	chemical compound space
EO	electro-optic
LCAP	linear combination of atomic potentials
NLO	nonlinear optical
nm	the official abbreviation for *nanometer* from the International System of Units
TD-DFT	time-dependent density-functional theory
VP-DFT	variation-of-particles density-functional-theory

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
1 (PDF)	DIR USARL RDRL WMM G J ORLICKI